Wisniewska, N. & Mora, J. C. (2018). Pronunciation learning through captioned videos. In J. Levis (Ed.), *Proceedings of the 9th Pronunciation in Second Language Learning and Teaching conference*, ISSN 2380-9566, University of Utah, September, 2017 (pp. 204-215). Ames, IA: Iowa State University.

PRONUNCIATION LEARNING TROUGH CAPTIONED VIDEOS

Natalia Wisniewska, Universitat de Barcelona

Joan C. Mora, University of Barcelona

The current study investigates L2 learners' skills at integrating auditory and orthographic input while reading dynamic texts in L2-captioned video, as part of a broader research project investigating the role of exposure to L2-captioned video in L2 pronunciation development. Within this broader research goal, the eye movements of L1-Catalan/Spanish learners of L2-English (*n*=38) were recorded while watching short L2-captioned video clips. The Reading Index for Dynamic Text (Kruger & Steyn, 2013) was used as a measure of learners' amount of text processing, and an index of text-sound integration was computed by calculating the extent to which fixations on selected words synchronized with their auditory onsets. We also explored learners' individual differences in text-sound integration through a novel task that required learners to uncover text-sound mismatches. In addition, we measured learners' L2 segmentations skills through a word-spotting task (McQueen, 1996) and L2 proficiency through an Elicited Imitation Task (Ortega et al., 2002). The results shed light on the relationship between reading and audio-text integration skills, suggesting that efficient reading might be what leads to modality integration.

INTRODUCTION

Unlike other aspects of foreign language teaching (e.g. grammar, vocabulary), pronunciation does not often receive enough attention in second language (L2) classroom settings. It has been argued that pronunciation instruction had become a "casualty of Communicative Language Teaching" (Thomson & Derwing, 2015: 326). The problem with the pedagogical implementation of pronunciation instruction is twofold. First, it is not easy to embed a focus on pronunciation within a communicative teaching approach that emphasizes interaction (Mora & Levkina, 2017). Second, when pronunciation training takes place, it typically does so at an individual level, provided the learner has an interest and the willingness to seek out-of-the-classroom sources of input. One common source of exposure to L2 spoken input is watching movies in the L2.

Research has shown the pedagogical benefits of multimedia learning and the use of audio-visual materials (Mayer, 2009), such as subtitled video (Danan, 2004). Studies investigating the pedagogical potential of multimodality report improvement in listening comprehension (e.g., Vanderplank, 1988) and L2 vocabulary acquisition (e.g., Montero Pérez, Van Den Noortgate & Desmet, 2013). Such gains are explained by the notion of bimodal reinforcement in Paivio's (1986) Dual Coding Theory, which claims that the dual processing of auditory and visual information helps create and strengthen the mental representations of perceived objects, and therefore, promote learning. Could subtitled videos also enhance L2 pronunciation development?

Speakers of English as a foreign language at all levels of proficiency experience difficulty understanding words in a continuous stream of speech. For example, Charles and Trenkic (2015) asked L2 speakers in long-term-immersion to perform a shadowing task (Mitterer & McQueen, 2009) where they were asked to repeat back audio excerpts from TV programs. They found that L2 speakers failed to repeat about 30% of what they heard. This may be due to the challenging task of making use of L2-specific word boundary cues to identify words

(McQueen & Cutler, 1998) or to a likely mismatch between the incoming auditory input and the inaccurate phonological representations of L2 words in the learners' mental lexicon (Broersma, 2012). It is therefore important to investigate what sources of input exposure may have a positive effect on learners' improvement in L2 speech processing and comprehension.

Bimodal input in the form of subtitles can benefit pronunciation development in at least two ways. At the perceptual level, subtitles can aid the decoding and segmentation of speech by helping listeners map auditory input to linguistic form in running speech. The simultaneous presentation of word forms in written and auditory modalities facilitates word identification and lexical access and can boost speech processing (Mitterer & McQueen, 2009). We propose that efficiency in matching auditory and orthographic representations of words during extensive viewing would not only lead to gains in L2 speech processing, but it might also trigger changes in the phonological representation of words in the mental lexicon, eventually leading to more target-like pronunciation. Indirect evidence of the potential of bimodal input exposure for L2 phonological development is provided by Mitterer and McQueen (2009). In their study, Dutch participants exposed to English subtitles while watching videos in unfamiliar Scottish or Australian accents, outperformed participants who watched the videos without subtitles or with subtitles in the L1 on a shadowing task measuring L2 speech segmentation. They found that L2 subtitles enhanced foreign speech perception by boosting lexically-guided learning. Thus, extensive exposure to bimodal input provided by L2-captioned video may facilitate the development of more accurate L2 phono-lexical representations, eventually leading to improvement in L2 pronunciation.

The current study is a part of a larger project on the development of L2 pronunciation through L2 captioned videos, where the overall aim is to investigate the potential benefits of multimodality for pronunciation learning. In the current study, we explored L2 learners' ability to integrate auditory and textual input (i.e., spoken and written word forms) while reading dynamic texts (i.e., using subtitles) in L2-captioned video by temporally relating learners' eye fixations on words to their auditory onset. Previous research using eye-tracking measures to capture on-screen reading behavior suggests subtitled text benefits listening comprehension (Montero Perez *et al.,* 2013) and the incidental acquisition of L2 vocabulary (Bisson *et al*., 2014), but no research to date has examined the role of text-audio synchronization during bimodal input exposure.

Our working assumption is that more effective integration skills are likely to enhance gains in L2 speech perception and production over time via exposure to bimodal input. In addition, we assessed the contribution of individual differences in speech processing skills to L2 learners' processing and integration of auditory and textual input in dynamic texts. We addressed the following research questions:

1. To what extent can L2 learners integrate text and audio during bimodal input exposure through L2-captioned video?
2. Do individual differences in speech processing skills affect the effectiveness of text and audio processing during bimodal input exposure through L2-captioned video?

In order to address these questions it is crucial to assess the learners' L2 proficiency level. Given that L2 speech and text processing is less efficient in the L2 than it is in the L1 (Segalowitz, 2010), watching L2 captioned video is a particularly challenging activity for the L2 learner (especially at low proficiency levels) from a cognitive load perspective (Mayer, 2009). This is because multimodal processing requires resolving competition and integration of different sources of L2 input without having control over the speed of the information flow: the auditory input (the soundtrack), the action in the background (the scene), and the textual input (the subtitles). Under such circumstances, the L2 proficiency of learners as well as the use of their cognitive resources and speech processing skills are likely to contribute substantially to their language processing efficiency and consequently to how much they can benefit from exposure to L2 captioned video for L2 pronunciation development.

METHODOLOGY

In order to assess L2 learners' processing of dynamic texts, we tracked their eye movements while watching short excerpts from the TV series *Sherlock* (2010) subtitled in English. The selected excerpts (1.5 minutes) featured quiet indoor conversations between characters without action in the background. In addition, we assessed L2 learners' individual differences in speech processing skills by obtaining accuracy measures from a speech processing test battery that included tasks on L2 speech segmentation, statistical learning of sound sequences, and modality integration.

Participants

Thirty-eight L1-Spanish/Catalan learners of English participated in the study for course credit. We assessed their proficiency level in English through an elicited imitation task (Ortega et al., 2002) consisting of 30 sentences with high-frequency vocabulary items, ranging 7-19 syllables in length, and of increasing grammatical complexity (see appendix). Participants heard each sentence only once through headphones and were asked to repeat the sentences as accurately as they could after a 2-second delay signaled by a beep sound. Following the scoring rubric of Ortega et al. (2002) available in the IRIS digital repository (Marsden, Mackey, & Plonsky, 2016), the first author scored each sentence assigning 0, 1, 2, 3 or 4 points, depending on how much of the sentence could be repeated and the kind of errors produced (if any), to a maximum score of 120 for the 30 sentences presented. The scores obtained ranged from 30 to 118 (*M*=96.87; *SD*=18.86), suggesting a relatively advanced level of proficiency at the high end of the scores.

Eye-tracking measures

Participants' eye movements were tracked and recorded on a Tobii T120 eye-tracker while they watched seven short clips. They were told they would do a language comprehension task that required watching the clips carefully in order to answer a true/false comprehension question appearing on the screen at the end of each one of the clips. The questions, which posed no difficulty to participants and were expected to be answered correctly, served to maintain students' attention and engagement while watching. For the current study, the eye-gaze data corresponding to two of the seven clips (Table 1) were used in the computation of the eye-tracking measures, as we estimated these data to be sufficient to provide individual measures of reading behavior.

Table 1

Characteristics of the clips

Clip	Duration (sec)	Subtitles (*n*)	Words (*n*)	Subtitle length (words)	Word length (characters)
1	89	33	220	6.3	5.3
2	95	35	271	7.9	4.9

Reading Index of Dynamic Text (RIDT)

The RIDT assesses the amount of visual processing of dynamic text. It provides a reliable 0-1 index of the visual processing of text in subtitles (Kruger & Stein, 2013). It is based on the number of fixations (points where the eye rests while reading) on each word in every caption while penalizing for skipping, re-fixations and regressions (Figure 1). The more text a viewer processes, the larger the index score, so that a viewer fixating all the words in all captions would obtain a score approaching 1. We used the RIDT to determine how much text in the captions learners had processed.

$$\text{RIDT} = \frac{\text{number of unique fixations for } p \text{ in } s}{\text{number of standard words in } s} \times \frac{\text{average forward saccade length for } p \text{ in } s}{\text{standard word length for } v}$$

(*s*=subtitle; *p*=participant; *v*=video)

Figure 1. Reading Index for Dynamic Text (RIDT) formula

Audio-Text Synchronization

The extent to which a fixation on a word in a caption is synchronized with its auditory presentation may determine how effectively the viewer can map auditory input to linguistic form. This mapping may be crucial in pronunciation learning in that it may promote changes in the phonological form of the learners' L2 lexical representations.

In order to estimate the extent to which auditory and visual word forms are synchronized while watching captioned video we determined whether participants fixated on the written form of a selected set of target words either *before* or *after* the occurrence of its auditory form. We also calculated the time distance in milliseconds between the onset of the fixation on the written word and the onset of the auditory form of the word in the sound track. Ten target words were selected on the basis of their length and position in the sentence. Words shorter than five characters and those appearing at the beginning or at the end of the sentence in the caption were avoided, as they are often skipped, according to research on caption reading (d'Ydewalle & De Bruycker, 2007)**.**

Speech processing measures

In what follows we describe the tasks used to obtain measures of participants' individual differences in speech processing skills. We obtained measures of speech segmentation (word-spotting and statistical learning) and audio-text integration (modality integration tasks). The elicited imitation task (described in the participants section) provided a measure of L2 proficiency.

Word-spotting

Participants' L2 segmentation skills were assessed through a word-spotting task (Cutler & Shanley, 2010; McQueen, 1996). Participants identified real English words embedded in nonwords by saying them out loud. We asked a male and a female native speaker of British English to produce 72 nonwords from the nonword set created by Farrell (2015), which were recorded in a sound-proof booth. The position of the real word in the nonwords were *final* (*n*=36, e.g. *ke+song*=kesong) or *initial* (*n*=36, e.g. *pound+fisp=pound*fisp). Nonwords conformed to one of three conditions: (1) *Stress*: The nonsense syllable was unstressed (*n*=24, e.g. *map*ef); (2) *Easy*: the segmentation affected an illegal phonotactic sequence in English (*n*=24, e.g. *ink*fab) or (3) *Difficult*: a legal one in English (*n*=24, e.g. *step*lut). Nonwords were presented auditorily only and participants had 3.5 seconds to say the embedded word, which was recorded to compute a percent correct accuracy score. Learners with higher word-spotting scores were assumed to have better segmentation skills, which would provide them with an advantage in processing multimodal input.

Statistical learning

Individual differences in learners' ability to extract phonotactic regularities from sound sequences, an ability that supports the segmentation of words from fluent speech (Saffran, Aslin & Newport, 1996), were assessed through the statistical learning task in Palmer and Mattys (2016; stream A), based on Saffran et al. (1996). Participants were first exposed to a continuous stream of speech (6 minutes) consisting of 4 alien words (*laso*kachu, re*bufi*, *pema*dovi, ti*nugo*) and 4 part-words made up of word-final and word-initial syllables (*bufilaso*, *dovire*, *nugopema*, *kachuti*), each presented 125 times at a speech rate of 4.17 syllables per second. There were no acoustic segmentation cues within the speech stream, so that the relative between-syllable transitional probabilities within and across words was the main cue participants could use for inferring word boundaries. Then they performed a 24-trial recognition test where they were asked to identify which of two "alien" words, a possible one (one of the words or part-words) and a "nonword" which could not be made up of word-final and word-initial syllables (e.g., *finukado*) was a word in the alien language. Word pairs were presented auditorily and orthographically (500-ms ISI, 1000-ms ITI). We calculated an overall percent correct score based on the number of correctly identified words and part-words.

Modality integration tasks

Two modality integration tasks were designed to assess learners' skill at integrating text and sound in sentence-like contexts. Both tasks required learners to uncover text-sound mismatches as they simultaneously listened to and read short sentences appearing on the screen, half of which contained a text-sound mismatch. Their task was to decide whether what they heard matched what they read by pressing a designated *same* or *different* key on the keyboard. One task was in English (participants' L2) and the other one in a language participants could not understand (Basque). Thus, whereas in the English task participants had to process spoken and written word forms and their meaning, in the Basque task the processing of spoken and written word forms was based on phonetic and orthographic decoding only.

The English task contained two blocks of 24 sentence trials each produced by a female native speaker of Standard Southern British English (with 5 practice trials). In block A, the written sentences (adapted from Kennedy & Trofimovich, 2008) were presented in standard form. In block B, they were presented without spaces. In each block, half of the sentences were *same* trials (12) and half were *different* trials (12), and half of each one of these (6) contained a target nonword with an orthographic representation that either matched or mismatched its auditory form. We computed a percent correct identification score based on the number of correctly identified *same* and *different* trials.

The Basque task contained a single block of 20 sentence trials presented orthographically with spaces and produced by a female native speaker of Basque (with 4 practice trials), half of which (10) were *different* trials with text-sound mismatches. Learners who were better able to identify text-audio mismatches were assumed to show better text-audio integration skills, which would allow them to process multimodal input more effectively.

RESULTS AND DISCUSSION

Eye-tracking measures

The analysis of the eye-tracking data revealed a wide range of RIDT scores (Table 2), suggesting that the amount of text participants processed in the captions varied considerably across learners (from 0.16 to 0.78 within the 0-1 RIDT index). We were expecting the RIDT score to correlate negatively with learners' proficiency, i.e. higher proficiency learners were expected to skip more words and captions than lower proficiency learners, as they would need to rely less on the supporting text in the captions to understand spoken language. In general, irrespective of L2 proficiency, we expected learners with more efficient speech processing skills to experience less difficulty in managing different sources of information and in integrating text and sound more efficiently.

Table 2

Eye-tracking measures (n= 38)

Measure	*M*	*SD*	*Min.*	*Max.*
RIDT	0.52	0.18	0.16	0.78
% words pre-fixated	70.4	17.2	33.3	100
% words post-fixated	29.6	17.2	0	66.7
Pre-fixation distance (milliseconds)	216	198	0	842
Post-fixation distance (milliseconds)	531	185	108	1183
Fixation distance (milliseconds)	441	120	186	760

The text-sound synchronization measures revealed that most fixations (70.4%) occurred before the auditory presentation of the word (Figure 2), with a large number of participants (71.3%) mainly fixating on the selected words before they were presented auditorily. This indicates, as expected, that participants had generally already read the selected target words by the time these occurred in the soundtrack. Such pre-fixations took place on average about 200 milliseconds from the onset of the auditory word form, a text-to-sound distance that was significantly shorter (Wilkoxon: *T*=680, *z*=4.48, *p*<.001) than the average time distance of about 500 milliseconds between the auditory word form and the following post-fixation (Figure 2). For the set of selected words in this analysis, the tendency was therefore for the auditory word form to follow (rather than precede) the lexical activation of the word via the visual text input.

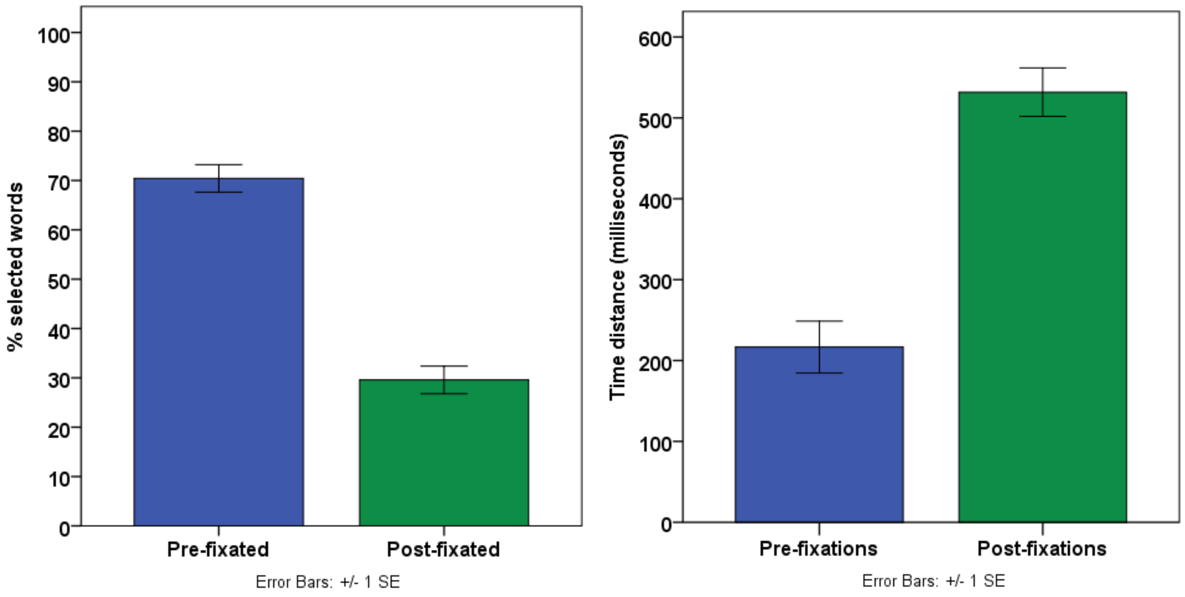

Figure 2. Percentage of selected words that were fixated before (pre-fixated) or after (post-fixated) their auditory onset (left) and distance in time between the eye-fixation on the word in the caption and its auditory onset (right).

Speech processing measures

The results obtained for the speech processing tasks (Table 3) also revealed large variability in L2 learners' performance both for tasks conducted in the L2 (reflecting inter-learner differences in L2 proficiency) and tasks conducted in a language unfamiliar to participants (L0).

Table 3

Descriptive data (% correct) for the speech processing tasks

Task	*M*	*SD*	*Min.*	*Max.*
Elicited imitation (L2)	56.7	9.7	41.7	81.9
Word-spotting (L2)	56.7	9.7	41.7	81.9
Statistical learning (L0)	56.6	15.9	25	96
English modality integration (L2)	81.6	9.4	55	95
Basque modality integration (L0)	75.1	9.4	55	95

The results of the elicited imitation task suggest an average upper intermediate level of proficiency, with scores ranging from intermediate to advanced proficiency levels.

Segmentation skills measured through the word spotting task, showed, as expected, differences in performance as a function of the *position* (initial, final) of the target word and *stimuli set* (stress, easy, difficult; see Figure 3). A two-way ANOVA with *position* and *stimuli set* as within-subjects factors revealed a non-significant main effect of *position* (*F*(1, 36)=.230, *p*=.634, η^2=006), a significant main effect of stimuli set (*F*(2, 35)=20.01, *p*<.001, η^2=533) and a significant *position* x *stimuli set* interaction (*F*(2, 35)=36.33, *p*<.001, η^2=675). As shown in Figure 2, the interaction arose because it was significantly harder to identify the target word in initial than in final position in the *stress* condition (*t*(36)=-5.61, *p*<.001), whereas the opposite happened in the *difficult* condition (*t*(36)=3.98, *p*<.001). The main effect of condition, however, was significant for both *initial* (*F*(2, 35)=13.77, *p*<.001, η^2=440) and *final* (*F*(2, 35)=27.99, *p*<.001, η^2=615) positions. In order to explore the relationship between segmentation skills and the eye-tracking measures (Table 3), we computed an average segmentation skill score across conditions by subject.

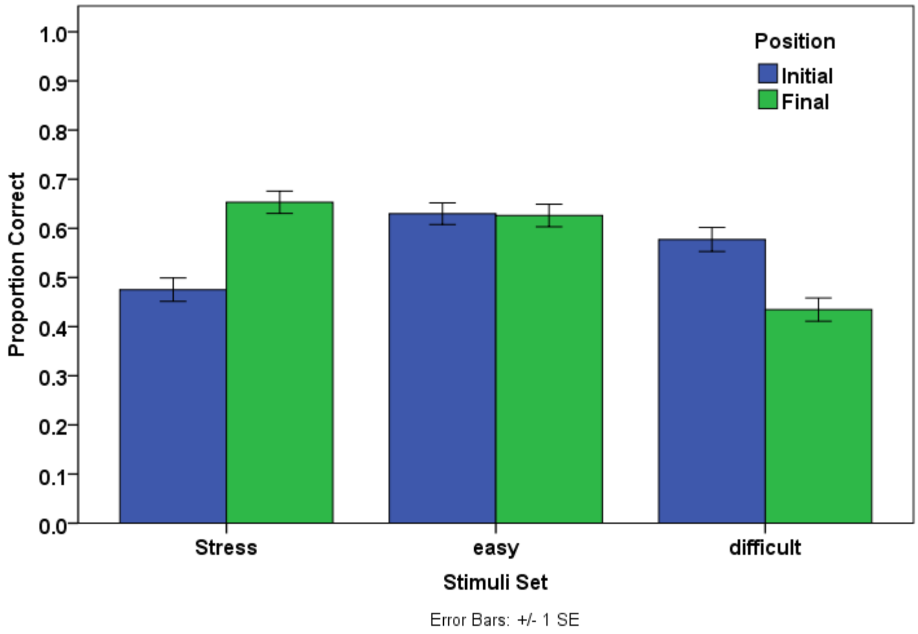

Figure 3. Proportion of correctly identified words by position and stimuli set.

The testing phase in the statistical learning task involved presenting participants with three sets of pairs of "alien" words for identification: *words* vs. *partwords*, *words* vs. *nonwords*, and *partwords* vs. *nonwords*. They were expected to identify *words* and *partwords* as words in the "alien" language at higher frequency rates than *nonwords*. As shown in Figure 4, mean scores by condition ranged between 50% and 60% (see Table 3). However, mean scores across testing conditions ranged from 25% to 96%, indicating large inter-learner variation, with 26 out of the 36 participants obtaining scores ≥ 50%. We computed a statistical learning score (Table 3) based on the average across the three testing conditions by subject.

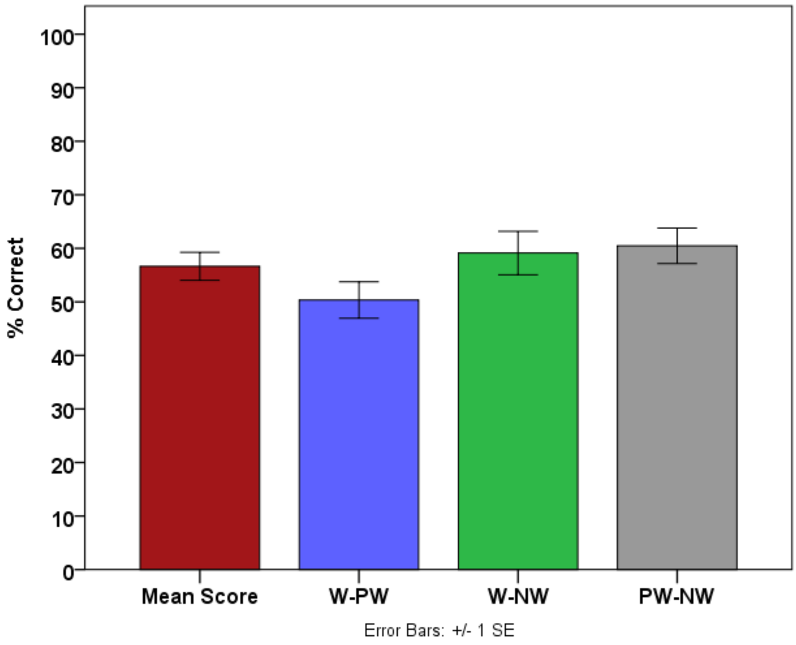

Figure 4. Proportion of correctly identified words (W) and partwords (PW).

Finally, as shown in Table 3, L2 learners' performance on the modality integration tasks also presented large inter-learner variation. Participants obtained significantly higher correct identification scores in the English than they did in the Basque version of the task (*t*(37)=-3.42, *p*=.002), suggesting that it was easier for them to identify mismatches when meaning processing was involved. Scores on both tasks appeared to be unrelated (*Pearson-r*=.141, *p*=.383), suggesting that they were tapping on different types of modality integration.

Correlation analyses

The main aim of the current study was to assess L2 learners' ability to integrate text and audio during bimodal input exposure through L2 caption video and to explore the role of individual differences in speech processing skills at doing so effectively.

We first explored whether our L2 proficiency measure was related to learners' performance in the L2 speech processing tasks. As expected, higher proficiency learners were better able to identify target words in the word-spotting task (*r*=.446, *p*=006), suggesting they could segment L2 speech better. They were also better able to detect text-sound mismatches in the English modality integration task (*r*=.513, *p*<.001), but not in the Basque modality integration task (*r*=-.030, *p*=856), suggesting that they were also better able to integrate text and sound in the L2. Proficiency scores, however, were found to be unrelated to any of the eye-tracking measures. Similarly, the L2 speech processing measures (word-spotting and English modality integration scores) were found to be unrelated to eye-tracking measures when L2 proficiency was controlled for. L0 speech processing measures (statistical learning and English modality integration scores) were unrelated to eye-tracking measures. However, as expected, higher RIDT indices were associated with the percentage of target words learners fixated on (*r*=.552, *p*<.001), suggesting that the RIDT index could reliably capture the amount of text learners processed in the captions. Finally, learners who were better able to identify text-sound mismatches in the Basque modality integration task had shorter time spans between auditory word onsets and fixations in general (*r*=-.327, *p*=.045), but larger time spans in post-fixations (*r*=.502, *p*=.001). This may be due to the fact that only a small proportion of fixations occurred after the auditory presentation of the word (only 29.6% were post-fixations), or it may indicate that text-sound synchronization skills are driven by the ability to read fast, suggesting that only "fast readers" can synchronize text and sound well.

CONCLUSION

This study was a first attempt at exploring the role of speech processing skills in bimodal input processing within our wider research aim (pronunciation development through L2 captioned videos). The results underscore the role of L2 proficiency in L2 speech processing and shed light on the mechanisms underlying text-sound synchronization during bimodal input processing in captioned video. Learners with higher L2 proficiency showed better L2 segmentation skills. However, whereas such skills were unrelated to eye-tracking measures, better L0 text-sound integration skills were related to longer time spans between auditory input and post-fixations, but not between auditory input and pre-fixations. This may suggest an important role for individual differences in reading speed (which we did not test) in text-sound integration when watching captioned video. Testing L1 and L2 silent reading speed and eye-gaze behavior is needed in order to better understand the mechanisms underlying text-sound integration during exposure to bimodal input and further explore their role in L2 pronunciation development. Our next step in this research project is to relate the measures explored in the current study to gains in L2 pronunciation obtained after extensive exposure to L2-captioned videos.

ACKNOWLEDGMENTS

We would like to thank Ingrid Mora-Plaza and Diana Moreira de Oliveira for their help in data collection.

ABOUT THE AUTHORS

Joan C. Mora is associate professor in the Department of Modern Languages and Literatures and English Studies in University of Barcelona (UB). His research has examined the role of contextual and individual factors in the development of L2 speech and oral fluency, and the acquisition of L2 phonology.

Contact information: mora@ub.edu, https://joancmora.weebly.com/

Natalia Wisniewska is a PhD candidate and research assistant collaborating with the Language Acquisition Research Group (GRAL) at the University of Barcelona. Her PhD dissertation focuses on the potential benefits of multimodal input for pronunciation development and the role of individual cognitive differences in the acquisition of L2 speech.

Contact information: wisniewska@ub.edu

REFERENCES

Bisson, M.J., Van Heuven, W.J., Conklin, K. and Tunney, R.J. (2014) Processing of native and foreign language subtitles in films: An eye tracking study. *Applied Psycholinguistics,* 35 (2), 399–418.

Broersma, M. (2012). Increased lexical activation and reduced competition in second-language listening. *Language and Cognitive Processes*, 27, 1205-1224.

Charles, T. J., & Trenkic, D. (2015). Speech segmentation in a second language: The role of bimodal input. In Y. Gambier, A. Caimi, & C. Mariotti (Eds.), *Subtitles and Language Learning: Principles, Strategies and Practical Experiences,* (pp. 173-198). Bern: Peter Lang.

Danan, M. (2004). Captioning and subtitling, undervalued language learning strategies. *Meta*, *49*(1), 67–77.

d'Ydewalle, G., & De Bruycker, W. (2007). Eye movements of children and adults while reading television subtitles. *European Psychologist, 12*(3), 196-205. DOI: 10.1027/10169040.12.3.196.

Kennedy, S., & Trofimovich, P. (2008). Intelligibility, comprehensibility, and accentedness of L2 speech: The role of listener experience and semantic context. *Canadian Modern Language Review*, *64*(3), 459-489.

Kruger, J. L., & Steyn, F. (2014). Subtitles and eye tracking: Reading and performance. *Reading Research Quarterly*, *49*(1), 105-120.

Marsden, E., Mackey A., & Plonsky, L. (2016). The IRIS Repository: Advancing research practice and methodology. In A. Mackey & E. Marsden (Eds.), *Advancing methodology and practice: The IRIS Repository of Instruments for Research into Second Languages* (pp. 1-21). New York: Routledge.

Mayer, R. E., (2009). *Multimedia Learning*. Second Edition. New York: Cambridge University Press.

McQueen, J. (1996). Word spotting. *Language and Cognitive Processes*, *11*(6), 695-699.

McQueen, J. M., & Cutler, A. (1998). Spotting (different kinds of) words in (different kinds of) context. In R. Mannell, & J. Robert-Ribes (Eds.), *Proceedings of the Fifth International Conference on Spoken Language Processing: Vol. 6* (pp. 2791-2794). Sydney: ICSLP.

Mitterer, H., & McQueen, J. M. (2009). Foreign subtitles help but native-language subtitles harm foreign speech perception. *PloS one*,*4*(11), 1-5.

Montero Perez, M., Van Den Noortgate, W., & Desmet, P. (2013). Captioned video for L2 listening and vocabulary learning: A meta-analysis. *System*, 41(3), 720-739.

Mora, J. C., Levkina, M., (2017). Task-based pronunciation teaching and research: key issues and future directions. *Studies in Second language Acquisition*, 39, 381-399.

Munro, M. J., Derwing, T. M., & Thomson, R. I. (2015). Setting segmental priorities for English learners: Evidence from a longitudinal study. *International Review of Applied Linguistics in Language Teaching*, *53*(1), 39-60.

Ortega, L., Iwashita, N., Norris, J. M., & Rabie, S. (2002, October). An investigation of elicited imitation tasks in crosslinguistic SLA research. Presentation given at *Second Language Research Forum,* Toronto.

Paivio, A. (1986). *Mental representations, A dual coding approach*. Oxford: Oxford University Press.

Palmer, S. D., & Mattys, S. L. (2016). Speech segmentation by statistical learning is supported by domain-general processes within working memory. *The Quarterly Journal of Experimental Psychology, 69*(12), 2390-2401.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, 1926-1928.

Segalowitz, N. (2010). *Cognitive Bases of Second Language Fluency*. New York: Routledge.

Sherlock (2010) Season 1, Episode 1: *A Study in Pink*, episode 2: *The Blind Banker*, TV programme. Directed by P. McGuigan. UK: BBC.

Thomson, R. I., & Derwing, T. M. (2015). The effectiveness of L2 pronunciation instruction: A narrative review. *Applied Linguistics*,36(3), 326-344

Vanderplank, R. (1988). The value of teletext subtitles in language learning. *ELT Journal, 42*(4)*,* 272–281.

APPENDIX

The sentences in the elicited imitation task (from Ortega et al., 2002) available from IRIS (Marsden, Mackey, & Plonsky, 2016)

1. I have to get a haircut.

2. The red book is on the table.

3. The streets in this city are wide.

4. He takes a shower every morning.

5. What did you say you were doing today?

6. I doubt that he knows how to drive that well.

7. After dinner I had a ling, peaceful nap.

8. It is possible that it will rain tomorrow.

9. I enjoy movies which have a happy ending.

10. The houses are very nice but too expensive.

11. The little boy whose kitten died yesterday is sad.

12. That restaurant is supposed to have very good food.

13. I want a nice, big house in which my animals can live.

14. You really enjoy listening to country music, don't you?

15. She just finished painting the inside of her apartment.

16. Cross the street at the light and then just continue straight ahead.

17. The person I'm dating has a wonderful sense of humor.

18. She only orders meat dishes and never eats vegetables.

19. I wish the price of town houses would become affordable.

20. I hope it will get warmer sooner this year than it did last year.

21. A good friend of mine always takes care of my neighbor's three children.

22. The black cat that you fed yesterday was the one chased by the dog.

23. Before he can go outside, he has to finish cleaning his room.

24. The most fun I've ever had was when we went to the opera.

25. The terrible thief whom the police caught was very tall and thin.

26. Would you be so kind as to hand me the book which is on the table?

27. The number of people who smoke cigars is increasing every year.

28. I don't know if the 11:30 train has left the station yet.

29. The exam wasn't nearly as difficult as you told me it would be.

30. There are a lot of people who don't eat anything at all in the morning.